Available online at www.sciencedirect.com

SCIENCE DIRECT®

LINEAR ALGEBRA AND ITS APPLICATIONS

ELSEVIER

Linear Algebra and its Applications 366 (2003) 25–37

www.elsevier.com/locate/laa

Polynomial factorization through Toeplitz matrix computations

Dario A. Bini [a,*], Albrecht Böttcher [b]

[a] *Dipartimento di Matematica, Università di Pisa, 56127 Pisa, Italy*
[b] *Fakultät für Mathematik, TU Chemnitz, 09107 Chemnitz, Germany*

Received 20 April 2001; accepted 4 September 2002

Submitted by G. Heinig

Abstract

The problem of polynomial factorization is translated into the problem of constructing a Wiener–Hopf factorization, and three algorithms are designed for the solution of the latter problem. These algorithms are based on solving linear systems with large (but finite) circulant and Toeplitz matrices. The algorithms are of low complexity and, perhaps most importantly, they are extremely lucid. An upper bound for the condition number of the problem of polynomial factorization is given in terms of the condition number of a certain Toeplitz matrix.

AMS classification: Primary 12Y05; Secondary 12D05, 15A23, 47B35, 65H05

Keywords: Polynomial factorization; Wiener–Hopf factorization; Toeplitz matrix; Circulant matrix; Finite section method; Condition number

1. Introduction

Suppose we are given a polynomial

$$p(z) = p_0 + \cdots + p_{n-1}z^{n-1} + z^n \quad (p_0 \neq 0) \tag{1}$$

and we know that $m \geqslant 1$ zeros are located in $\{z \in \mathbf{C} : |z| < r < 1\}$ and that $n - m \geqslant 1$ zeros are contained in $\{z \in \mathbf{C} : |z| > R > 1\}$. We want to factorize $p(z)$ into a product $v(z)\ell(z)$ of two polynomials

* Corresponding author.
E-mail addresses: bini@dm.unipi.it (D.A. Bini), aboettch@mathematik.tu-chemnitz.de (A. Böttcher).

0024-3795/03/$ - see front matter
doi:10.1016/S0024-3795(02)00594-3

$$v(z) = v_0 + \cdots + v_{m-1}z^{m-1} + z^m, \tag{2}$$

$$\ell(z) = \ell_0 + \cdots + \ell_{n-m-1}z^{n-m-1} + z^{n-m} \tag{3}$$

such that all zeros of $v(z)$ (resp. $\ell(z)$) are of modulus less than r (resp. greater than R). Without loss of generality we assume that $m \geqslant n - m$.

Several algorithms for the solution of this problem were proposed in [11,15–17,19]. The algorithm designed by Pan [15,16] has the best asymptotic computational complexity bound, but its implementation is not straightforward due to its high logical complexity and, moreover, its effectiveness is still unclear. For the so-called problem of spectral factorization, that is, for the case where $n = 2m$ and $p_{n-j} = \bar{p}_j$ for $j = 0, 1, \ldots, m$, easy-to-use algorithms are described and analyzed in [8].

It is well known since about Gohberg and Feldman's book [7] that the above problem is equivalent to the construction of a Wiener–Hopf factorization of the Laurent polynomial $a(z) = z^{-m}p(z)$. However, except for the algorithm of [10] and the MinPh algorithm of [8], this equivalence has not been thoroughly exploited until the recent papers [1–3]. Paper [3] is a systematic treatise of the interplay between polynomial computations and their counterparts in Toeplitz matrices, and it contains a detailed analysis of iterative algorithms of the Graeffe type for obtaining a Wiener–Hopf factorization.

The purpose of this paper is to reveal some more of the benefits one can receive by passing to Toeplitz matrices. We show three extremely simple ways of translating the problem of polynomial factorization into problems on the inversion of infinite matrices. The latter problems are solved by the finite section method. What results is three algorithms that are distinguished by great lucidity and low complexity. Moreover, our approach allows us to establish bounds for the condition number of the problem of polynomial factorization and to give asymptotic bounds for the accuracy of our algorithms. This paper does not contain deep new theorems, and all the mathematics we are using is well known. However, we believe that the ease of all our arguments discloses the essence of the matter and shows the full beauty of the idea of replacing polynomial factorization by Toeplitz matrix computations, without any veiling by too many technical details.

We thank Bernd Silbermann for drawing our attention to the finite section method in connection with the problem of Wiener–Hopf factorization and the referees for valuable remarks on an earlier version of this paper.

2. Toeplitz matrices

Let $\mathbf{T}$ be the complex unit circle. Given a continuous function $c : \mathbf{T} \to \mathbf{C}$ with Fourier coefficients $\{c_k\}_{k\in\mathbf{Z}}$, we let

$$T(c) = (c_{j-k})_{j,k=1}^{\infty} \quad \text{and} \quad T_q(c) = (c_{j-k})_{j,k=1}^{q}$$

denote the infinite and $q \times q$ Toeplitz matrices generated by c. The matrix $T(c)$ induces a bounded operator on $\ell^2(\mathbf{N})$ whose norm is $\|c\|_\infty$, the maximum of $|c(z)|$

over $|z| = 1$. We think of $T_q(c)$ as an operator on $\mathbf{C}^q$ with the ℓ^2 norm. The operator $T(c)$ is invertible if and only if c has no zeros on $\mathbf{T}$ and the winding number of the map $c : \mathbf{T} \to \mathbf{C}$ is zero (see, e.g., [5,7]).

Now let $p(z)$, $v(z)$, $\ell(z)$ be the polynomials (1), (2), (3), put

$$a(z) = z^{-m}p(z) = p_0 z^{-m} + \cdots + p_{n-1}z^{n-m-1} + z^{n-m},$$
$$u(z) = z^{-m}v(z) = 1 + v_{m-1}z^{-1} + \cdots + v_0 z^{-m}, \tag{4}$$

and leave $\ell(z)$ as it is. The problem raised in the introduction is equivalent to the search for a Wiener–Hopf factorization, that is, for a representation $a(z) = u(z)\ell(z)$ where $\ell(z)$ and $u(z)$ are of the form (3) and (4), $\ell(z) \neq 0$ for $|z| \leqslant R$ and $u(z) \neq 0$ for $|z| \geqslant r$. It can be readily verified that if $a(z) = u(z)\ell(z)$ is a Wiener–Hopf factorization, then $T(u)$ is upper triangular, $T(\ell)$ is lower triangular, and

$$T(a) = T(u)T(\ell),$$
$$T^{-1}(u) = T(u^{-1}), \quad T_q^{-1}(u) = T_q(u^{-1}),$$
$$T^{-1}(\ell) = T(\ell^{-1}), \quad T_q^{-1}(\ell) = T_q(\ell^{-1}),$$

where $T^{-1}(u) := (T(u))^{-1}$, etc. The following result is well known.

Lemma 2.1. *If $q \geqslant n - m$, then $T_q(a^{-1})$ is an invertible matrix.*

Proof. There are several ways to prove this (see, e.g., [7, Lemma III.6.1] or [4,14]). The perhaps simplest proof is as follows. Let $L(c) = (c_{j-k})_{j,k=-\infty}^{\infty}$ be the Laurent matrix generated by a continuous function $c : \mathbf{T} \to \mathbf{C}$. The matrix $L(c)$ induces a bounded operator on $\ell^2(\mathbf{Z})$, and $L(c)$ is invertible if and only if c has no zeros on $\mathbf{T}$, in which case $L^{-1}(c) = L(c^{-1})$. Let R_q be the projection on $\ell^2(\mathbf{Z})$ defined by $(R_q y)_j = y_j$ for $j \in \{1, \ldots, q\}$ and $(R_q y)_j = 0$ otherwise. Put $S_q = I - R_q$. The operator

$$T_q(a^{-1}) = R_q L(a^{-1}) R_q|\mathrm{Im}\, R_q$$

is invertible if and only if $S_q L(a) S_q|\mathrm{Im}\, S_q$ is invertible (see, e.g., Lemma 2.9 of [5]). The matrix of the latter operator splits into four natural blocks. One of the non-diagonal blocks is zero, because $q \geqslant n - m$, and the diagonal blocks are the Toeplitz matrix generated by a and the transpose of this matrix. As these two matrices are invertible, so also is $S_q L(a) S_q|\mathrm{Im}\, S_q$. □

Let $x = T_q(\ell)e_1$ where $e_1 = (1\ 0\ 0\ \cdots)^{\mathrm{T}}$. Clearly, x gives us the first q coefficients $\ell_0, \ldots, \ell_{q-1}$ of $\ell(z)$ ($\ell_j := 0$ for $j > n - m$). The following simple result provides us with three identities for x that can be used to compute x.

Proposition 2.2. *We have*

$$T_q(a^{-1})x = e_1 \quad \text{for } q > n - m, \tag{5}$$

$$T_q(\ell^{-1})x = e_1 \quad \text{for } q \geqslant 1, \tag{6}$$

$$x = T_q(u^{-1})T_q(a)e_1 \quad \text{for } q > n - m. \tag{7}$$

Proof. Let P_q be the projection on $\ell^2(\mathbf{N})$ that acts by the rule $(P_q y)_j = y_j$ for $j \in \{1, \ldots, q\}$ and $(P_q y)_j = 0$ otherwise. Set $Q_q = I - P_q$. Then

$$\begin{aligned} T_q(a^{-1})T_q(\ell)e_1 &= P_q T(a^{-1})P_q T(\ell)e_1 \\ &= P_q T(a^{-1})T(\ell)e_1 - P_q T(a^{-1})Q_q T(\ell)e_1 \\ &= P_q T(a^{-1})T(\ell)e_1 \text{ (because } Q_q T(\ell)e_1 = 0 \text{ if } q > n - m) \\ &= P_q T(u^{-1})T(\ell^{-1})T(\ell)e_1 = P_q T(u^{-1})e_1 = e_1, \end{aligned}$$

which is (5). Equality (6) is trivial. Finally,

$$\begin{aligned} T_q(u^{-1})T_q(a)e_1 &= T_q(u^{-1})P_q T(u)T(\ell)e_1 \\ &= T_q(u^{-1})P_q T(u)P_q T(\ell)e_1 + T_q(u^{-1})P_q T(u)Q_q T(\ell)e_1 \\ &= T_q(u^{-1})P_q T(u)P_q T(\ell)e_1 \text{ (because } Q_q T(\ell)e_1 = 0 \text{ if } q > n - m) \\ &= T_q(u^{-1})T_q(u)P_q T(\ell)e_1 = P_q T(\ell)e_1 = x, \end{aligned}$$

which completes the proof of (7). □

Formula (5) immediately yields the following strategy: compute the Fourier coefficients of a^{-1} and solve the Toeplitz system $T_q(a^{-1})x = e_1$ (recall Lemma 2.1). To employ formula (6), we compute the first q entries of the first column of $T^{-1}(a)$. Since $T^{-1}(a)e_1 = T(\ell^{-1})T(u^{-1})e_1 = T(\ell^{-1})e_1$, these entries form just the first column of the Toeplitz matrix $T_q(\ell^{-1})$ and thus determine $T_q(\ell^{-1})$ completely. It remains to solve the lower-triangular Toeplitz system $T_q(\ell^{-1})x = e_1$. To take advantage of (7), we compute the first q entries $d_0, \ldots, d_{q-1}$ of the first row of $T^{-1}(a)$. Clearly, $e_1^{\mathrm{T}}T^{-1}(a) = e_1^{\mathrm{T}}T(\ell^{-1})T(u^{-1}) = \ell_0^{-1}e_1^{\mathrm{T}}T(u^{-1})$. As the 1, 1 entry of $T(u^{-1})$ is 1, it follows that $\ell_0 = d_0^{-1}$ and that the first row of $T_q(u^{-1})$ is $(1, d_1 d_0^{-1}, \ldots, d_{q-1}d_0^{-1})$. With the first row we have all of $T_q(u^{-1})$. We finally multiply the upper-triangular Toeplitz matrix $T_q(u^{-1})$ by $T_q(a)e_1$, that is, by the first column of $T_q(a)$.

Thus, what we are left with is algorithms for computing either the $2q - 1$ central Fourier coefficients of a^{-1}, or the first q entries of the first column of $T^{-1}(a)$, or the first q entries of the first row of $T^{-1}(a)$. Equivalently, we need $P_q T(a^{-1})P_q$ or the first column or first row of $P_q T^{-1}(a)P_q$.

3. Algorithms based on the finite section method

The first rows and columns of $P_q T^{-1}(a) P_q$ can be obtained by the finite section method: it is well known (see, e.g., [5,7]) that if a is continuous and $T(a)$ is invertible, then the matrices $T_N(a)$ are invertible for all sufficiently large N and $T_N^{-1}(a)P_N$ converges strongly to $T^{-1}(a)$ on $\ell^2(N)$. In particular,

$$P_q T_N^{-1}(a)e_1 \to P_q T^{-1}(a)e_1, \quad e_1^{\mathrm{T}} T_N^{-1}(a) P_q \to e_1^{\mathrm{T}} T^{-1}(a) P_q. \tag{8}$$

In order to execute the finite section method, we need to know something about the N's for which $T_N(a)$ is invertible and something about the rate of convergence in (8). These questions have been studied in [13,14,18], for example (also see [4,5,7]). In particular, the following result is known. We give a sketch of the proof for the reader's convenience. Put

$$\varrho = \max(r, 1/R), \quad \sigma = r/R.$$

Theorem 3.1. *There exist positive constants $\gamma(a, r, R)$ and $\delta(a, r, R)$ depending only on a, r, R such that if $\gamma(a, r, R)\sigma^N < 1$, then $T_N(a)$ is invertible and*

$$\left\| P_q T_N^{-1}(a)e_1 - P_q T^{-1}(a)e_1 \right\|_{\ell^1} \leqslant \delta(a, r, R)\varrho^N, \tag{9}$$

$$\left\| e_1^{\mathrm{T}} T_N^{-1}(a) P_q - e_1^{\mathrm{T}} T^{-1}(a) P_q \right\|_{\ell^1} \leqslant \delta(a, r, R)\varrho^N. \tag{10}$$

Proof. Let $K = T(a^{-1}) - T^{-1}(a)$. The operator $T_N(a) = P_N T(a) P_N|\mathrm{Im}\, P_N$ is invertible if and only if

$$Q_N T^{-1}(a) Q_N|\mathrm{Im}\, Q_N = Q_N(T(a^{-1}) - K)Q_N|\mathrm{Im}\, Q_N$$

is invertible (again see, e.g., Lemma 2.9 of [5]). The latter operator is certainly invertible if

$$\|Q_N K Q_N\| \, \|T^{-1}(a^{-1})\| < 1, \tag{11}$$

where here and in the following $\|\cdot\|$ is always the operator norm on ℓ^2. Standard computations (see, e.g., [5]) give

$$\begin{aligned} K &= H_+(\ell^{-1})H_-(u^{-1}) \qquad (12)\\ &= H_+(a^{-1}u)H_-(\ell a^{-1}) = H_+(a^{-1})[T(u)]^{\mathrm{T}}[T(\ell)]^{\mathrm{T}}H_-(a^{-1})\\ &= H_+(a^{-1})[T^{-1}(a^{-1})]^{\mathrm{T}}H_-(a^{-1}), \qquad (13)\end{aligned}$$

where $[\cdot]^{\mathrm{T}}$ denotes transposition and the Hankel operators $H_\pm(c)$ are defined by

$$H_+(c) = (c_{j+k-1})_{j,k=1}^{\infty}, \quad H_-(c) = (c_{-j-k+1})_{j,k=1}^{\infty}.$$

For $j \geqslant 1$,

$$(a^{-1})_j = \frac{1}{2\pi \mathrm{i}} \int_{|z|=1} \frac{z^{-j-1}\,\mathrm{d}z}{a(z)} = \frac{1}{2\pi \mathrm{i}} \int_{|z|=R} \frac{z^{-j-1}\,\mathrm{d}z}{a(z)},$$

whence $|(a^{-1})_j| \leqslant (\min_{|z|=R} |a(z)|)^{-1}(1/R)^j$. This implies that

$$\|Q_N H_+(a^{-1})\| \leqslant \sum_{j=N+1}^{\infty} |(a^{-1})_j| \leqslant \frac{1}{\min_{|z|=R} |a(z)|} \frac{1}{R-1} \frac{1}{R^N}. \tag{14}$$

Analogously one can show that

$$\|H_-(a^{-1})Q_N\| \leqslant \frac{1}{\min_{|z|=r} |a(z)|} \frac{1}{1/r-1} r^N. \tag{15}$$

Estimating $\|Q_N K Q_N\|$ with the help of (13)–(15) we see that (11) holds if

$$\frac{\|T^{-1}(a^{-1})\|^2}{\min_{|z|=r} |a(z)| \min_{|z|=R} |a(z)|} \frac{1}{R-1} \frac{1}{1/r-1} \left(\frac{r}{R}\right)^N < 1. \tag{16}$$

Using (12) instead of (13) we obtain in a similar way that (11) is valid whenever

$$\frac{\|T^{-1}(a^{-1})\|}{\min_{|z|=r} |u(z)| \min_{|z|=R} |\ell(z)|} \frac{1}{R-1} \frac{1}{1/r-1} \left(\frac{r}{R}\right)^N < 1. \tag{17}$$

Finally, proceeding exactly as in the proof of Theorem 2.15 of [5] (but considering all operators on the space ℓ^1), one can verify that

$$\left\| T_N^{-1}(a)e_1 - P_N T^{-1}(a)e_1 \right\|_{\ell^1} \leqslant \delta(a, r, R)\varrho^N,$$

where $\delta(a, r, R)$ is some constant. This gives (9). Estimate (10) can be proved analogously. □

Theorem 3.1 tells us that the first N_0 such that $T_N(a)$ is invertible for $N \geqslant N_0$ is generically not an astronomic number. Clearly, (16) and (17) provide us with different upper bounds for the constant $\gamma(a, r, R)$. In some cases things are actually much simpler. For instance, if the distance of the convex hull of the range $a(\mathbf{T})$ to the origin is positive (which happens, in particular, if a is positive), then $T_N(a)$ is invertible for all $N \geqslant 1$ by virtue of the Brown–Halmos theorem (see Proposition 2.17 of [5]).

Let now $T_q(\tilde{\ell}^{-1})$ and $T_q(\ell^{-1})$ be the lower-triangular Toeplitz matrices whose first columns are $P_q T_N^{-1}(a)e_1$ and $P_q T^{-1}(a)e_1$, respectively. By Theorem 3.1,

$$\begin{aligned} \left\| T_q(\tilde{\ell}^{-1}) - T_q(\ell^{-1}) \right\| &\leqslant \|\tilde{\ell}^{-1} - \ell^{-1}\|_\infty \leqslant \sum_{j=0}^{q-1} \left| (\tilde{\ell}^{-1})_j - (\ell^{-1})_j \right| \\ &= \left\| P_q T_N^{-1}(a)e_1 - P_q T^{-1}(a)e_1 \right\|_{\ell^1} = \mathrm{O}(\varrho^N). \end{aligned}$$

Thus, for the solutions of $T_q(\tilde{\ell}^{-1})\tilde{x} = e_1$ and $T_q(\ell^{-1})x = e_1$ we obtain

$$\frac{\|\tilde{x} - x\|_2}{\|x\|_2} \leqslant \frac{\|T_q(\ell)\|\mathrm{O}(\varrho^N)}{1 - \|T_q(\ell)\|\,\mathrm{O}(\varrho^N)} \leqslant \|\ell\|_\infty\,\mathrm{O}(\varrho^N) + \mathrm{O}(\varrho^{2N}). \tag{18}$$

Algorithm 1. We now arrive at the following algorithm with formula (6). Fix $q > n - m$ and choose $N \geqslant q$ so that ϱ^N is of the same order as the admitted relative approximation error $\varepsilon > 0$ and so that $T_N(a)$ is invertible. Of course, we may choose q and N as powers of 2. Solve the system $T_N(a)\tilde{y} = e_1$ (or find at least the first q components of the solution), which costs $\mathrm{O}(N \log^2 N)$ operations, let $T_q(\tilde{\ell}^{-1})$ be the lower-triangular Toeplitz matrix whose first column is $(\tilde{y}_0, \ldots, \tilde{y}_{q-1})^{\mathrm{T}}$, and solve the system $T_q(\tilde{\ell}^{-1})\tilde{x} = e_1$, which costs $\mathrm{O}(q \log q)$ operations. The resulting $\tilde{x}$ approximates the vector x of the coefficients of ℓ, and the error is $\|\tilde{x} - x\|_2/\|x\|_2 = \mathrm{O}(\varepsilon)$.

Algorithm 2. In a similar way we can exploit (7). Fix q and N as powers of 2 such that $N \geqslant q > n - m$, such that ϱ^N is of the same order as ε, and such that $T_N(a)$ is invertible. Solve the system $e_1^{\mathrm{T}} = \tilde{y}^{\mathrm{T}} T_N(a)$ (or find at least the first q components of the solution), let $T_q(\tilde{u}^{-1})$ be the upper-triangular Toeplitz matrix whose first row is $(1, \tilde{y}_1\tilde{y}_0^{-1}, \ldots, \tilde{y}_{q-1}\tilde{y}_0^{-1})$, and denote by $\tilde{x}$ the product of $T_q(\tilde{u}^{-1})$ and the first column of $T_q(a)$. The costs are $\mathrm{O}(N \log^2 N) + \mathrm{O}(q \log q)$. The $\tilde{x}$ obtained approximates the coefficients x of ℓ and

$$\frac{\|\tilde{x} - x\|_2}{\|x\|_2} \leqslant \ell_0\|u^{-1}\|_\infty \mathrm{O}(\varrho^N) + \mathrm{O}(\varrho^{2N}) = \mathrm{O}(\varepsilon).$$

We now turn to an algorithm based on (5). Such an algorithm requires the central $q \times q$ block of the Laurent matrix $L^{-1}(a) = L(h) = (h_{j-k})_{j,k=-\infty}^{\infty}$ generated by $h := a^{-1}$. The inverses of Laurent matrices can be approximated by the inverses of circulant matrices. Let $N > n$ be an even number (in practice it will be a power of 2) and let $C_N(a)$ be the $N \times N$ circulant matrix whose first row is

$$\begin{pmatrix} a_0 & a_{-1} & \cdots & a_{-m} & 0 & \cdots & 0 & a_{n-m} & a_{n-m-1} & \cdots & a_1 \end{pmatrix}.$$

Lemma 3.2. *The matrix of $C_N(a)$ is invertible for all $N > n$ and*

$$C_N^{-1}(a) = \left(\frac{1}{N} \sum_{\mu=0}^{N-1} \frac{\overline{\omega}_N^{\mu(j-1)} \omega_N^{\mu(k-1)}}{a(\omega_N^\mu)} \right)_{j,k=1}^{N}, \tag{19}$$

where $\omega_N = \exp(2\pi \mathrm{i}/N)$.

Proof. The circulant matrix with the first row $(b_0\ b_{N-1}\ b_{N-2}\ \cdots\ b_1)$ can be represented in the form

$$U_N^* \operatorname{diag}\Big(b(\omega_N^j)\Big)_{j=0}^{N-1} U_N, \tag{20}$$

where $U_N := \left(1/\sqrt{N}\right)\left(\omega_N^{(j-1)(k-1)}\right)_{j,k=1}^{N}$ and $b(z) := b_0 + b_1 z + \cdots + b_{N-1}z^{N-1}$. In our case

$$b(z) = a_0 + a_1 z + \cdots + a_{n-m}z^{n-m} + a_{-m}z^{N-m} + \cdots + a_{-1}z^{N-1},$$

and since $\omega_N^{jN} = 1$ for all j, it follows that $b(\omega_N^j) = a(\omega_N^j)$ for all j. As $a(z) \neq 0$ for $|z| = 1$, the matrix (20) has the inverse

$$U_N^* \operatorname{diag}\left(1/b(\omega_N^j)\right)_{j=0}^{N-1} U_N = U_N^* \operatorname{diag}\left(1/a(\omega_N^j)\right)_{j=0}^{N-1} U_N,$$

which gives (19). □

Lemma 3.2 implies that the j,k entry $[C_N^{-1}(a)]_{j,k}$ converges to

$$\frac{1}{2\pi}\int_0^{2\pi} \frac{e^{-\mathrm{i}\theta(j-k)}}{a(e^{\mathrm{i}\theta})}\mathrm{d}\theta = h_{j-k} = [L^{-1}(a)]_{j-k}.$$

In particular,

$$\tilde{h}_k(N) := \frac{1}{N}\sum_{\mu=0}^{N-1} \frac{\overline{\omega}_N^{\mu k}}{a\left(\omega_N^{\mu}\right)} \to h_k \quad \text{as } N \to \infty \tag{21}$$

for $k = -N/2, \ldots, N/2-1$. As the following theorem shows, the convergence in (21) is exponential.

Theorem 3.3. *There exists a positive constant $\eta(a, r, R)$ depending only on a, r, R such that $|\tilde{h}_k(N) - h_k| \leqslant \eta(a, r, R)\varrho^{N/2}$ for $k = -N/2, \ldots, N/2-1$.*

Proof. Define $g_N(j)$ by

$$h\left(\omega_N^j\right) = \sum_{k=-N/2}^{N/2-1} h_k \omega_N^{jk} + g_N(j).$$

It can be easily checked that

$$\tilde{h}_k(N) - h_k = \frac{1}{N}\sum_{j=0}^{N-1} g_N(j)\overline{\omega}_N^{jk}. \tag{22}$$

Estimating the Fourier coefficients of $h = a^{-1}$ as in the proof of Theorem 3.1, we get

$$|h_k| \leqslant \max\left(\frac{1}{\min_{|z|=r}|a(z)|}, \frac{1}{\min_{|z|=R}|a(z)|}\right)\varrho^{|k|},$$

whence $|g_N(j)| \leqslant \eta(a, r, R)\varrho^{N/2}$ for all j. The assertion is now obvious from equality (22). □

Put

$$\tilde{h}(z) = \sum_{k=-N/2}^{N/2-1} \tilde{h}_k(N) z^k \tag{23}$$

and let x and $\tilde{x}$ be the solutions of the systems $T_q(h)x = e_1$ and $T_q(\tilde{h})\tilde{x} = e_1$. We then have

$$\frac{\|\tilde{x} - x\|_2}{\|x\|_2} \leqslant \frac{\|T_q^{-1}(h)\| \, \|T_q(\tilde{h} - h)\|}{1 - \|T_q^{-1}(h)\| \, \|T_q(\tilde{h} - h)\|},$$

and since $T_q^{-1}(h) = T_q^{-1}(a^{-1})$ and

$$\|\tilde{h} - h\|_\infty \leqslant \sum_{k=-N/2}^{N/2-1} |\tilde{h}_k(N) - h_k| = \mathrm{O}(N\varrho^N),$$

we arrive at the estimate

$$\frac{\|\tilde{x} - x\|_2}{\|x\|_2} = \left\|T_q^{-1}(a^{-1})\right\| \mathrm{O}(N\varrho^N) + \mathrm{O}(N^2\varrho^{2N}).$$

Algorithm 3. In summary, formula (5) leads to the following algorithm. Choose q and N as powers of 2 such that $N \geqslant q > n - m$ and such that $N\varrho^N$ is comparable with ε. Compute

$$\tilde{h}_k(N) := \frac{1}{N} \sum_{\mu=0}^{N-1} \frac{\overline{\omega}_N^{\mu k}}{a\left(\omega_N^{\mu}\right)},$$

which can be done with $\mathrm{O}(N \log N)$ operations. Define $\tilde{h}$ by (23) and solve the system $T_q(\tilde{h})\tilde{x} = e_1$, which costs $\mathrm{O}(q \log^2 q)$ operations. The resulting vector $\tilde{x}$ is an approximation of the coefficients x of ℓ such that $\|\tilde{x} - x\|_2/\|x\|_2 = \mathrm{O}(\varepsilon)$.

Of course, part of Algorithm 3 resembles the evaluation/interpolation technique proposed in [6], where the power sums

$$\sum_{j=1}^{m} \zeta_j^k = \frac{1}{2\pi \mathrm{i}} \int_{\mathbf{T}} \frac{z^k p'(z) \, \mathrm{d}z}{p(z)}$$

of the zeros ζ_j of $p(z)$ in the open unit disk are computed by numerically integrating $z^k p'(z)/p(z)$ along the unit circle. The coefficients of the factor $u(z)$ are recovered from the power sums by solving Newton's equations.

Different algorithms for approximating the central $q \times q$ block of the Laurent matrix $L^{-1}(a)$ and for approximating the first q entries of $T^{-1}(a)e_1$ are described in [3]. These algorithms, which have doubly exponential convergence, are based

on Graeffe iterations and cyclic reduction. The acceleration provided by the doubly exponential convergence is paid by a higher logical complexity of the algorithms and by a computational cost per step which, even though asymptotically the same, involves larger multiplicative constants. Thus, the problem of selecting the most suitable algorithm for an effective implementation of polynomial factorization remains delicate. Clearly, the overall performance of each algorithm depends on the features of the specific polynomial to be factored.

4. The condition number of polynomial factorization

Using Toeplitz matrices, we can estimate the condition number of the problem of polynomial factorization very easily. Let $p(z)$ and $\tilde{p}(z)$ be two polynomials of the form (1) and suppose $\|\tilde{p} - p\|_\infty \leqslant \varepsilon \|p\|_\infty$. If $\varepsilon > 0$ is sufficiently small, then $\tilde{p}(z) = \tilde{v}(z)\tilde{\ell}(z)$ with polynomials $\tilde{v}(z)$ and $\tilde{\ell}(z)$ of the form (2) and (3) such that all zeros of $\tilde{v}(z)$ (resp. $\tilde{\ell}(z)$) have modulus less than r (resp. larger than R). Put $a(z) = z^{-m}p(z)$ and let $\|\cdot\|_2$ be the norm in $L^2(\mathbf{T})$; thus, if $c(z) = \sum_j c_j z^j$ is a polynomial, then $\|c\|_2^2 = 2\pi \sum_j |c_j|^2$.

Proposition 4.1. *For all sufficiently small $\varepsilon > 0$,*

$$\frac{\|\tilde{\ell} - \ell\|_2}{\|\ell\|_2} \leqslant \varepsilon \operatorname{cond} T(a^{-1}) \frac{\max|p(z)|}{\min|p(z)|} + \mathrm{O}(\varepsilon^2),$$

where $\operatorname{cond} T(a^{-1}) := \|T(a^{-1})\|\|T^{-1}(a^{-1})\|$ *and the maximum and minimum are taken over $z \in \mathbf{T}$.*

Proof. Let $\tilde{a}(z) = z^{-m}\tilde{p}(z)$. Then $\tilde{a} = a + \mu$ with $\|\mu\|_\infty \leqslant \varepsilon\|a\|_\infty$. We know from Proposition 2.2 that the coefficients of ℓ and $\tilde{\ell}$ are the solutions x and $\tilde{x}$ of the systems $T(a^{-1})x = e_1$ and $T(\tilde{a}^{-1})\tilde{x} = e_1$. Clearly, $\tilde{a}^{-1} = a^{-1} + \sigma$ with $\sigma = -\mu a^{-1}/(a + \mu)$. It follows that

$$\frac{\|\tilde{\ell} - \ell\|_2}{\|\ell\|_2} = \frac{\|\tilde{x} - x\|_2}{\|x\|_2} \leqslant \frac{\|T^{-1}(a^{-1})\|\,\|T(\sigma)\|}{1 - \|T^{-1}(a^{-1})\|\,\|T(\sigma)\|},$$

and since

$$\begin{aligned}\|T(\sigma)\| \leqslant \|\sigma\|_\infty &\leqslant \frac{\|\mu\|_\infty\|a^{-1}\|_\infty}{\min|a(z) + \mu(z)|} \leqslant \frac{\varepsilon\|a\|_\infty\|a^{-1}\|_\infty}{\min|a(z)| - \varepsilon\|a\|_\infty} \\ &= \frac{\varepsilon\|a\|_\infty\|a^{-1}\|_\infty}{\min|a(z)|} + \mathrm{O}(\varepsilon^2) = \varepsilon\|T(a^{-1})\|\frac{\max|p(z)|}{\min|p(z)|} + \mathrm{O}(\varepsilon^2),\end{aligned}$$

we arrive at the assertion. □

In the case of spectral factorization, Proposition 4.1 can be simplified to the following.

Corollary 4.2. *If $n = 2m$ and $p_{n-j} = \bar{p}_j$ for $j = 0, 1, \ldots, m$, then*

$$\frac{\|\tilde{\ell} - \ell\|_2}{\|\ell\|_2} \leqslant \varepsilon \left(\frac{\max |p(z)|}{\min |p(z)|} \right)^2 + \mathrm{O}(\varepsilon^2)$$

for all sufficiently small $\varepsilon > 0$, the maximum and minimum over $|z| = 1$.

Proof. The conditions imposed upon p imply that $a(z) = z^{-m} p(z)$ is real valued for $z \in \mathbf{T}$. As a has no zeros on $\mathbf{T}$, the function a is either positive or negative on $\mathbf{T}$. But in these cases $\operatorname{cond} T(a^{-1}) = \max |a| / \min |a|$, so that the corollary is immediate from Proposition 4.1. □

The following result provides us with an a posteriori estimate for the condition number of $T(a^{-1})$, which plays an important role in Theorem 3.1 and Proposition 4.1.

Proposition 4.3. *Let $\tilde{\ell}$ and $\tilde{u}$ be polynomials of the form* (3) *and* (4) *such that $\tilde{\ell}(z) \neq 0$ for $|z| \leqslant 1$ and $\tilde{u}(z) \neq 0$ for $1 \leqslant |z| \leqslant \infty$. Suppose $a = \tilde{u}\tilde{\ell} + b$ and $|b(z)| \leqslant \varepsilon |\tilde{u}(z)\tilde{\ell}(z)|$ for $z \in \mathbf{T}$, where ε is some constant satisfying $0 < \varepsilon < 1/2$. Then*

$$\operatorname{cond} T(a^{-1}) \leqslant \frac{1-\varepsilon}{1-2\varepsilon} \|a^{-1}\|_\infty \|\tilde{u}\|_\infty \|\tilde{\ell}\|_\infty.$$

Proof. Put $c := \tilde{\ell}^{-1} b \tilde{u}^{-1}$ and $(1+c)^{-1} =: 1 + d$. We then have the factorization $T(a^{-1}) = T(\tilde{u}^{-1}) T(1+d) T(\tilde{\ell}^{-1})$ (see, e.g., [5, Proposition 1.13]). Since $\|c\|_\infty \leqslant \varepsilon < 1/2$, it follows that $\|d\|_\infty \leqslant \varepsilon/(1-\varepsilon) < 1$. This implies that $T(1+d) = I + T(d)$ is invertible and that

$$\|T^{-1}(1+d)\| \leqslant \frac{1}{1 - \|d\|_\infty} \leqslant \frac{1-\varepsilon}{1-2\varepsilon}.$$

As $T^{-1}(a^{-1}) = T(\tilde{\ell}) T^{-1}(1+d) T(\tilde{u})$, we obtain that

$$\|T(a^{-1})\| \|T^{-1}(a^{-1})\| \leqslant \|a^{-1}\|_\infty \|\tilde{\ell}\|_\infty \frac{1-\varepsilon}{1-2\varepsilon} \|\tilde{u}\|_\infty. \qquad \square$$

The numerical stability of Algorithms 1–3 is easily understood. These algorithms can be divided into two stages: the approximation of a finite portion of the solution of an infinite system and the subsequent solution of a finite Toeplitz system. For instance, the first stage of Algorithm 3 consists in approximating $2q - 1$ coefficients of a^{-1}, i.e., the central entries of the Laurent matrix $L^{-1}(a)$. The second stage requires the solution of the system $T_q(a^{-1})x = e_1$. The condition number of the

former computation is bounded by $\operatorname{cond} L(a) = \max|a(x)|/\min|a(x)|$, while the condition number of the latter is $\operatorname{cond} T_q(a^{-1})$. Therefore both stages are numerically well conditioned if the number $\operatorname{cond} T_q(a^{-1}) \max|a(z)|/\min|a(z)|$ is not too large. Notice that $\operatorname{cond} T_q(a^{-1})$ converges to $\operatorname{cond} T(a^{-1})$ as $q \to \infty$ (see Corollary 3.9 of [5]). Thus, Algorithm 3 is numerically stable if the two stages are solved with numerically stable algorithms. The same cannot be said, for example, of the (quite popular) algorithms which are based on König's theorem, where in the second stage we encounter immensely large condition numbers.

König's theorem (see [9, Theorem 3.1.2] and [12]) says the following. Let $N > m+1$, let $T_N(p)$ be the $N \times N$ lower-triangular Toeplitz matrix whose first column is $(p_0 \;\; p_1 \;\; \cdots \;\; p_{N-1})^{\mathrm{T}}$, where $p_k := 0$ for $k > n$, and let $T_N(c)$ be the inverse of $T_N(p)$. Notice that $T_N(c)$ is also a lower-triangular Toeplitz matrix. Let V_N denote the lower-left $(m+1) \times (m+1)$ block of $T_N(c)$. Then V_N is invertible for all sufficiently large N, and if $x = (x_0 \;\; x_1 \;\; \cdots \;\; x_m)^{\mathrm{T}}$ is the solution of $V_N x = e_1$, then $x_j/x_m = v_j + \mathrm{O}((r/R)^N)$ for $j = 0, 1, \ldots, m-1$, where $v_0, v_1, \ldots, v_{m-1}$ are the coefficients of polynomial (2). This theorem reduces polynomial factorization to computing $m+1$ components of a large lower-triangular Toeplitz system and to solving an $(m+1) \times (m+1)$ Toeplitz system. Both computations can be performed with a low computational cost by means of FFT. However, one can show that $\operatorname{cond} V_N \to \infty$ as $N \to \infty$, which reveals that any algorithm based on König's theorem is numerically unstable if only N has to be chosen very large. The following simple example illustrates this feature.

The polynomial $p(z) = \sum_{i=0}^{10} z^i + 4z^5$ has five zeros of modulus less than 0.83 and five zeros of modulus greater then 1.21. Hence, we may take $r/R = 0.68$. Since $p(z)$ is positive and $1.5 < p(z) \leqslant 15$ for $|z| = 1$, the coefficient $(\max p(z)/\min p(z))^2$ in Corollary 4.2 is less than 100. Table 1 shows the condition numbers of V_N and of $T_8(\tilde{h})$, where $\tilde{h}$ is computed through (19) and (21). Table 2 reports the maximum modulus of the coefficients of the residual error $\tilde{v}(z)\tilde{\ell}(z) - p(z)$, where the approximate factors $\tilde{v}(z)$ and $\tilde{\ell}(z)$ are determined by the algorithm based on König's theorem and by Algorithm 3. The computation was performed with

Table 1
Condition numbers of the matrices V_N and $T_8(\tilde{h})$

N	16	32	64	128	256
$\operatorname{cond} V_N$	2.8×10^4	1.6×10^8	1.5×10^{16}	7.6×10^{32}	1.9×10^{64}
$\operatorname{cond} T_8(\tilde{h})$	4.2	3.6	3.6	3.6	3.6

Table 2
Residual errors generated by the algorithm based on König's theorem and by Algorithm 3

N	16	32	64	128	256
König's	1.6×10^{-2}	2.0×10^{-5}	5.0×10^{-11}	5.0×10^{-7}	3.2
Algorithm 3	7.8×10^{-1}	3.7×10^{-2}	6.0×10^{-5}	6.7×10^{-11}	8.8×10^{-16}

Mathematica^{TM}, where the default precision of 16 decimal digits was used for the computation of the residual errors and the larger precision of 100 decimal digits for the computation of the spectral condition number of V_N. Table 2 might suggest that König's theorem works very well, because the result for $N = 32$ or $N = 64$ is satisfactory and much better than the quality gotten with Algorithm 3. However, in the case at hand the polynomial is of the modest degree $n = 10$, which is not yet the terrain where instabilities may cause havoc.

References

[1] D.A. Bini, Using FFT-based techniques in polynomial and matrix computations: recent advances and applications, Numer. Funct. Anal. Optim. 21 (2000) 47–66.
[2] D.A. Bini, L. Gemignani, B. Meini, Factorization of analytic functions by means of Koenig's theorem and Toeplitz computations, Numer. Math. 89 (1) (2001) 49–82.
[3] D.A. Bini, L. Gemignani, B. Meini, Computations with infinite Toeplitz matrices and polynomials, Linear Algebra Appl. 343–344 (2002) 21–61.
[4] A. Böttcher, B. Silbermann, Notes on the asymptotic behavior of block Toeplitz matrices and determinants, Math. Nachr. 98 (1980) 183–210.
[5] A. Böttcher, B. Silbermann, Introduction to Large Truncated Toeplitz Matrices, Universitext, Springer-Verlag, New York, 1999.
[6] L.M. Delves, J.N. Lyness, A numerical method for locating the zeros of an analytic function, Math. Comput. 21 (1967) 543–560.
[7] I. Gohberg, I.A. Feldman, Convolution Equations and Projection Methods for Their Solution, AMS, Providence, RI, 1974 (Trans. Mathematical Monographs, vol. 41).
[8] T.N.T. Goodman, C.A. Micchelli, G. Rodriguez, S. Seatzu, Spectral factorization of Laurent polynomials, Adv. Comput. Math. 7 (1997) 429–454.
[9] A.S. Householder, The Numerical Treatment of a Single Nonlinear Equation, in: International Series in Pure and Applied Mathematics, McGraw-Hill, New York, 1970.
[10] D. Kershaw, An analysis of the method of L. Fox and L. Hayes for the factorization of a polynomial, Linear Algebra Appl. 86 (1987) 179–187.
[11] P. Kirrinnis, Partial fraction decomposition in $\mathbf{C}(z)$ and simultaneous Newton iteration for factorization in $\mathbf{C}[z]$, J. Complexity 14 (1998) 378–444.
[12] J. König, Über eine Eigenschaft der Potenzreihen, Math. Ann. 23 (1884) 447–449.
[13] S. Levin, Asymptotic properties of Toeplitz matrices, Dissertation, Weizmann Institute of Sciences, Rehovot, 1980.
[14] S. Levin, On invertibility of finite sections of Toeplitz matrices, Applicable Anal. 13 (1982) 173–184.
[15] V.Y. Pan, Optimal (up to polylog factors) sequential and parallel algorithms for approximating complex polynomial zeros, in: Proceedings of the 27th Annual ACM Symposium on Theory of Computing, ACM Press, New York, 1995, pp. 741–750.
[16] V.Y. Pan, Optimal and nearly optimal algorithms for approximating polynomial zeros, Comput. Math. Appl. 31 (1996) 97–138.
[17] V.Y. Pan, Solving a polynomial equation: some history and recent progress, SIAM Rev. 39 (1997) 187–220.
[18] A. Pomp, Normabschätzungen für die Inversen von Toeplitz–Matrizen, Z. Anal. Anwendungen 2 (1983) 175–187.
[19] A. Schönhage, The Fundamental Theorem of Algebra in Terms of Computational Complexity, University of Tübingen, Tübingen, 1982.